Visual Computing &
Artificial Intelligence
Prof. Matthias Nießner

Revisiting Structure from Motion with 3D Reconstruction Priors

Guided Research WS24/25

Daniel Korth
Advisor: Prof. Matthias Nießner

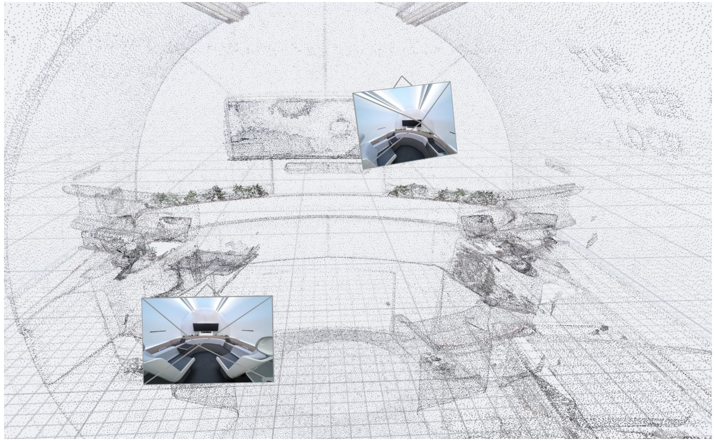
30.05.2025

Introduction

Structure from Motion

2D Image Collection

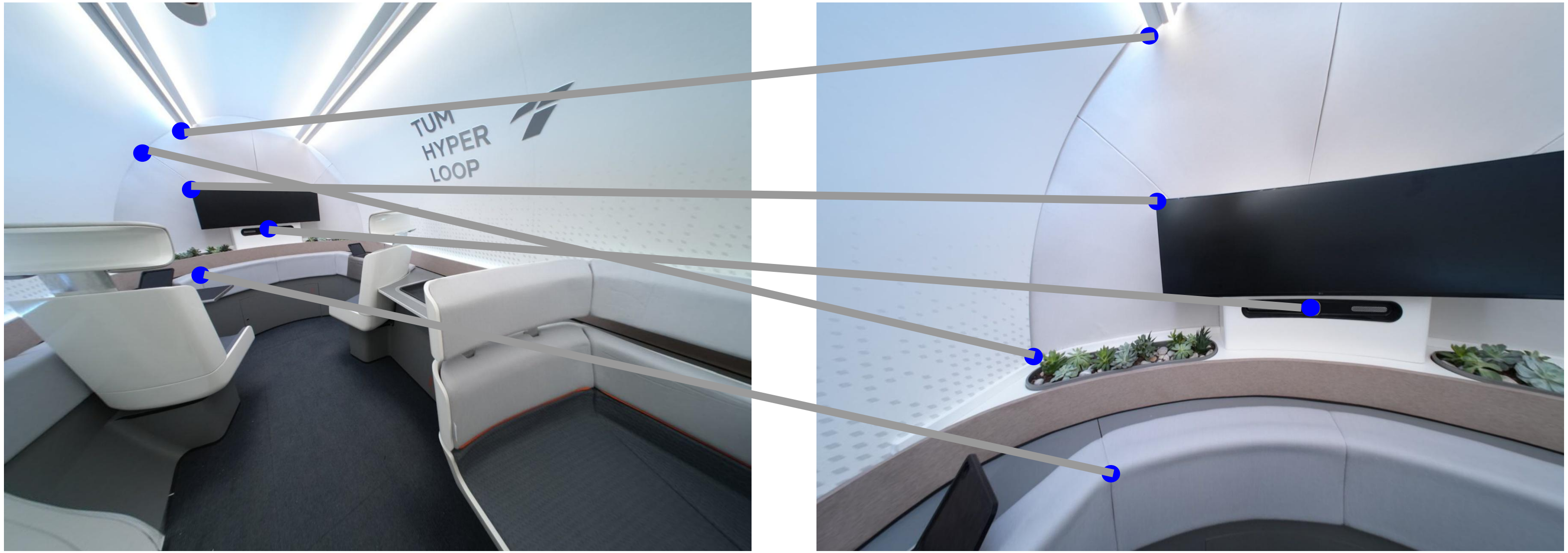
3D Reconstruction



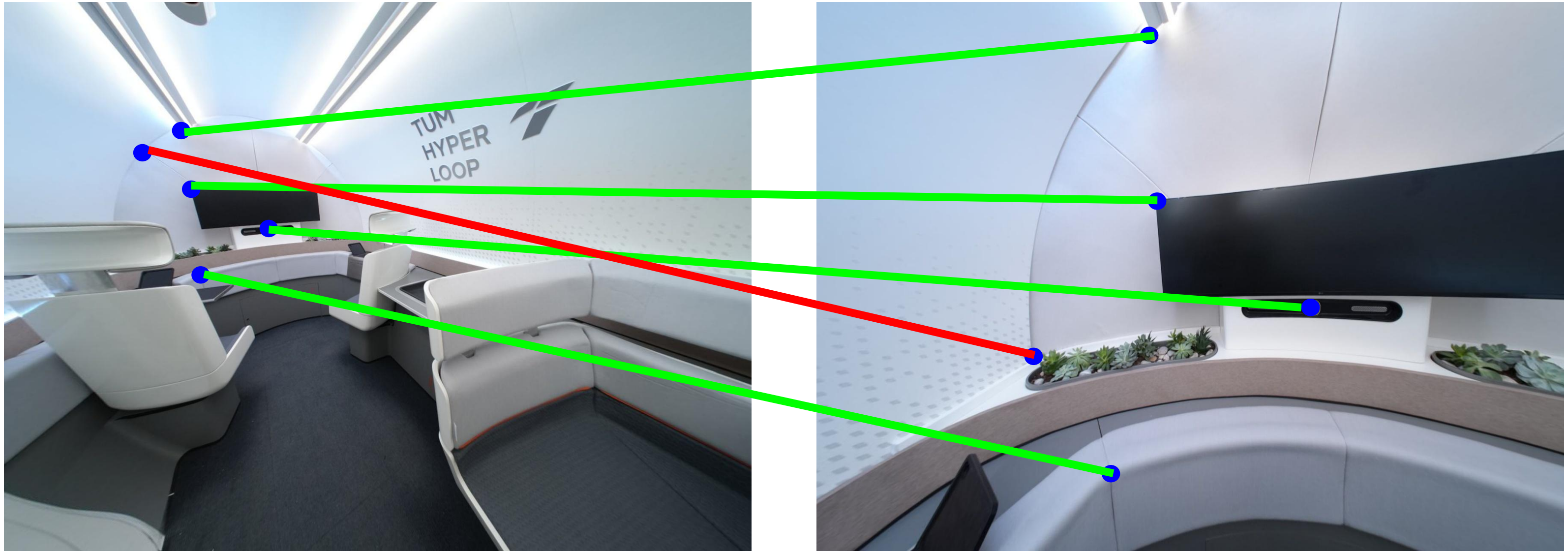
Camera Position + 3D Points

Detection + Description

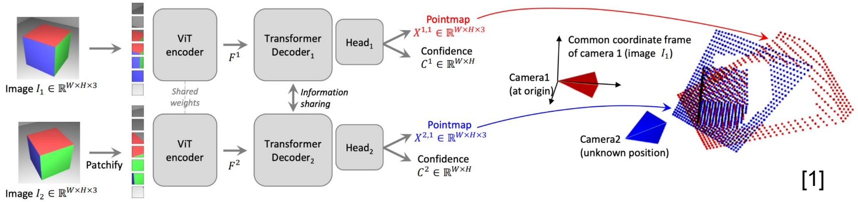
Descriptor Matching



Geometric Verification



3D Reconstruction Networks



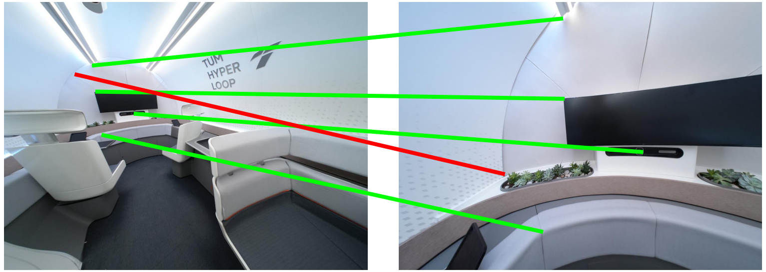
[1]

[1] Wang et al., “Dust3r: Geometric 3d vision made easy”, CVPR 2024

SfM Optimization

Global Optimization

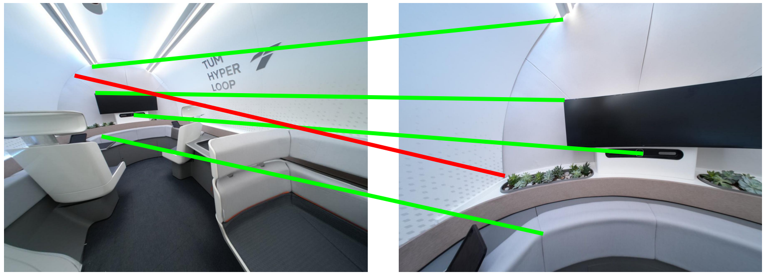
2D Constraints



SfM Optimization

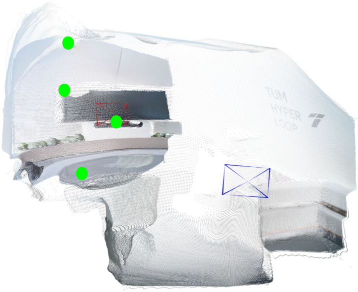
Global Optimization

2D Constraints



3D Recon Networks

dense 2D-3D mapping

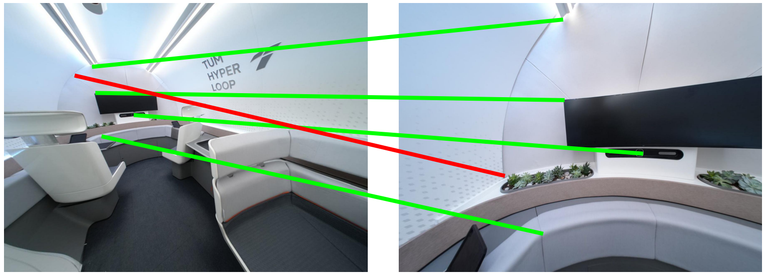


SfM Optimization

Global Optimization

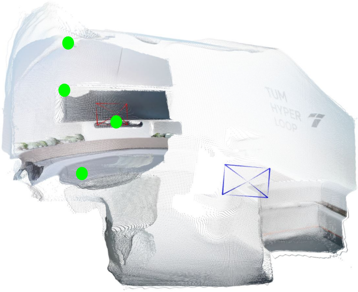
2D Constraints

?
3D Constraints



3D Recon Networks

dense 2D-3D mapping

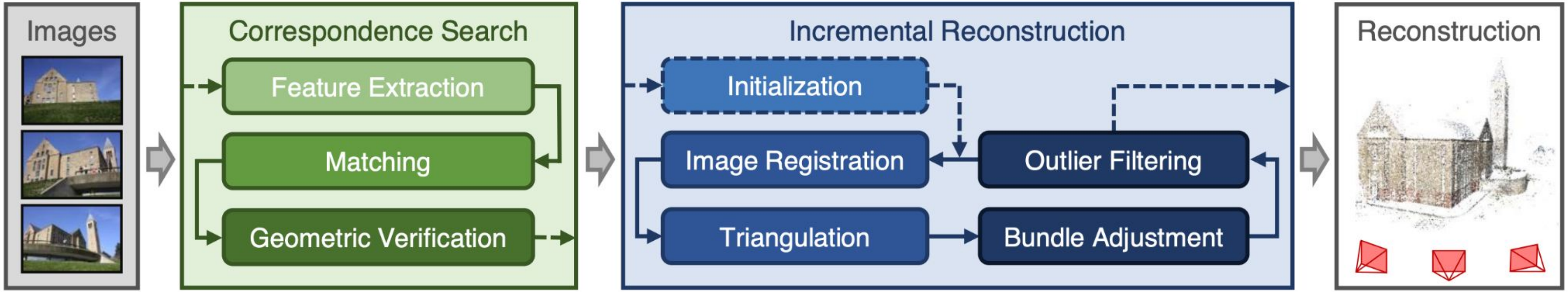


Goal

Add 3D Constraints to SfM Pipeline

Related Work

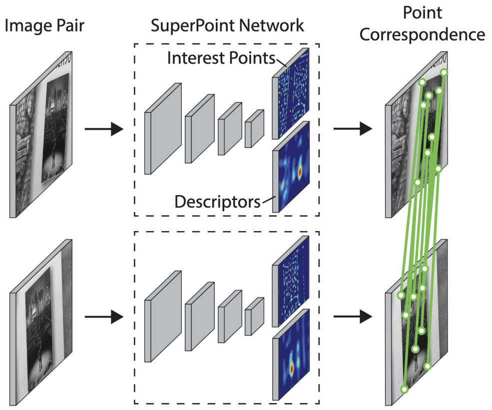
Incremental SfM Pipeline



[1] Schönberger and Frahm, "Structure-from-motion revisited", CVPR 2016

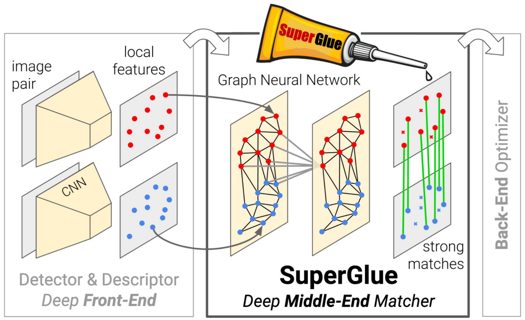
Modern SfM

Detection / Description



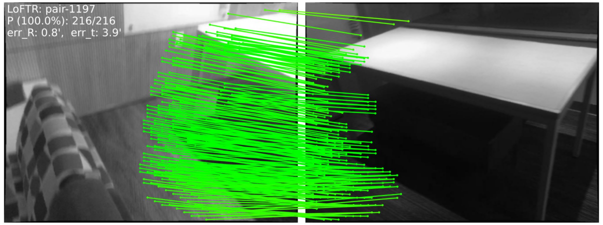
SuperPoint [1],
DeDoDe [2], …

Descriptor Matching



SuperGlue [3],
LightGlue [4], …

Dense Matching



LoFTR [5],
RoMA [6], …

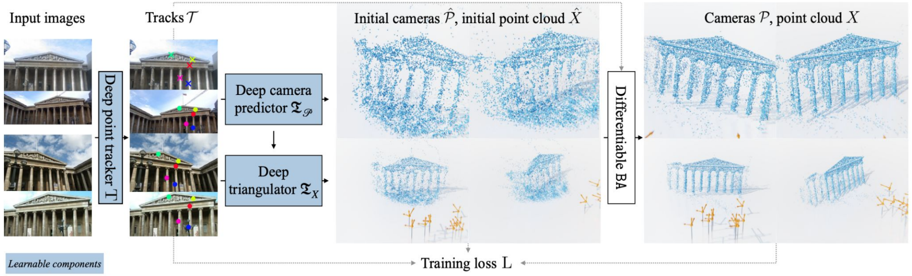
[1] DeTone et al., “SuperPoint: Self-supervised interest point detection and description”, CVPRW 2018
[2] Edstedt et al., “DeDoDe: Detect, Don’t Describe - Describe, Don’t Detect for Local Feature Matching”, 3DV 2024ç

[3] Sarlin et al., “SuperGlue: Learning feature matching with graph neural networks”, CVPR 2020
[4] Lindenberger et al., “LightGlue: Local Feature Matching at Light Speed”, ICCV 2023

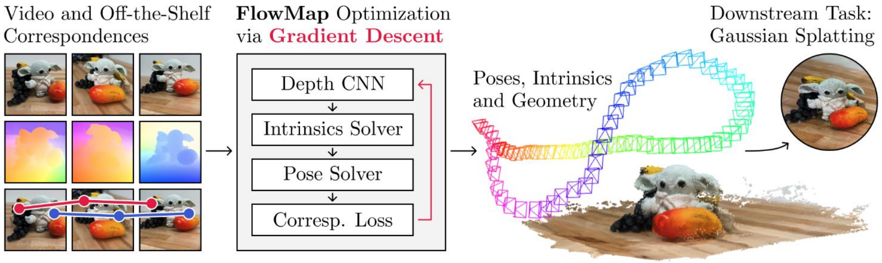
[5] Sun et al., “LoFTR: Detector-free local feature matching with transformers”, CVPR 2021
[6] Edstedt et al., “RoMA: Robust Dense Feature Matching”, CVPR 2024

End-to-End differentiable SfM

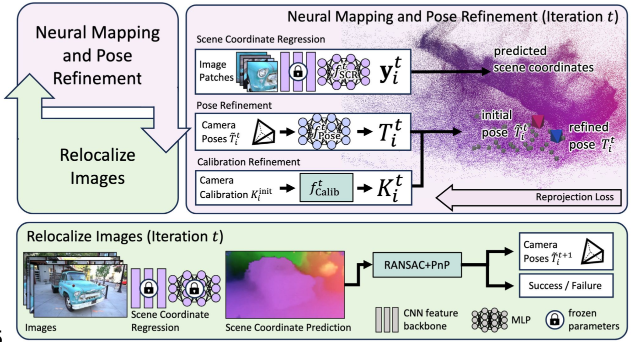
VGGSfM [1]



FlowMap [2]



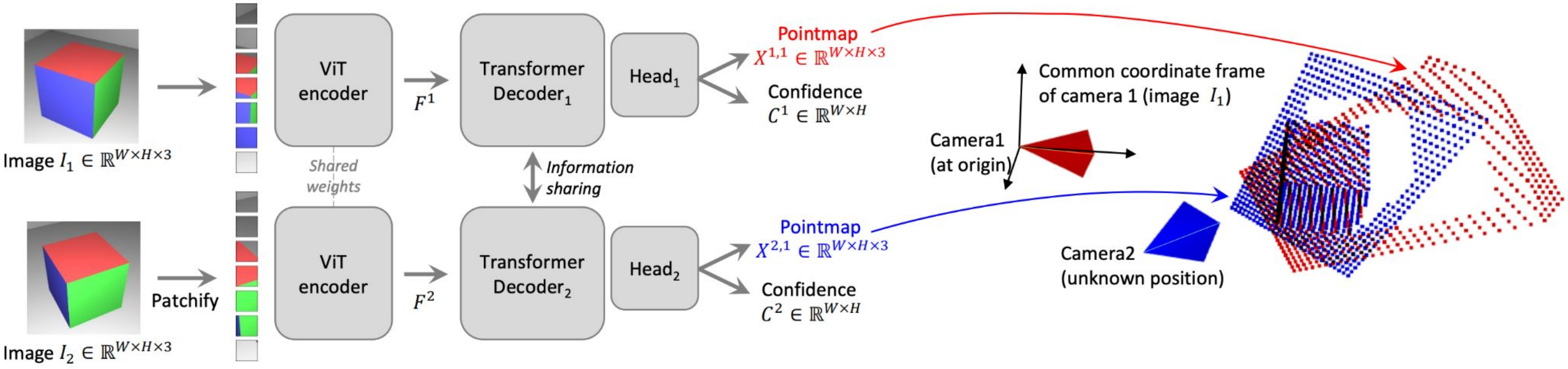
ACE0 [3]



[1] Wang et al., "VGGSfM: Visual geometry grounded deep structure from motion", CVPR 2024
[2] Smith & Charatan et al., "Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent", 3DV 2025
[3] Brachmann et al., "Scene Coordinate Reconstruction: Posing of image collections via incremental learning of a relocalizer", ECCV 2024

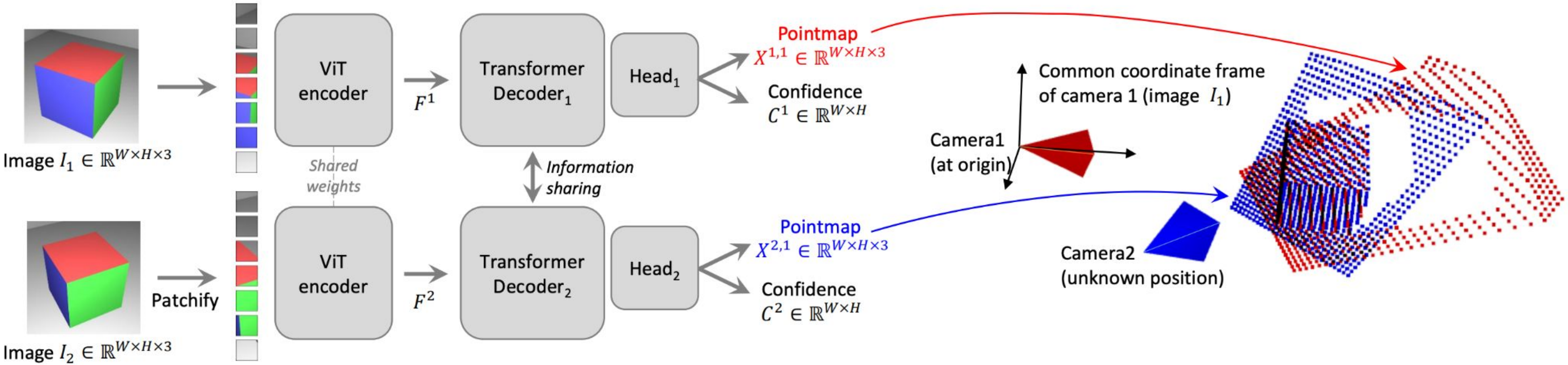
3D Reconstruction Networks - DUSt3R [1]

[1] Wang et al., "Dust3r: Geometric 3d vision made easy", CVPR 2024



3D Reconstruction Networks - DUSt3R [1]

[1] Wang et al., "Dust3r: Geometric 3d vision made easy", CVPR 2024

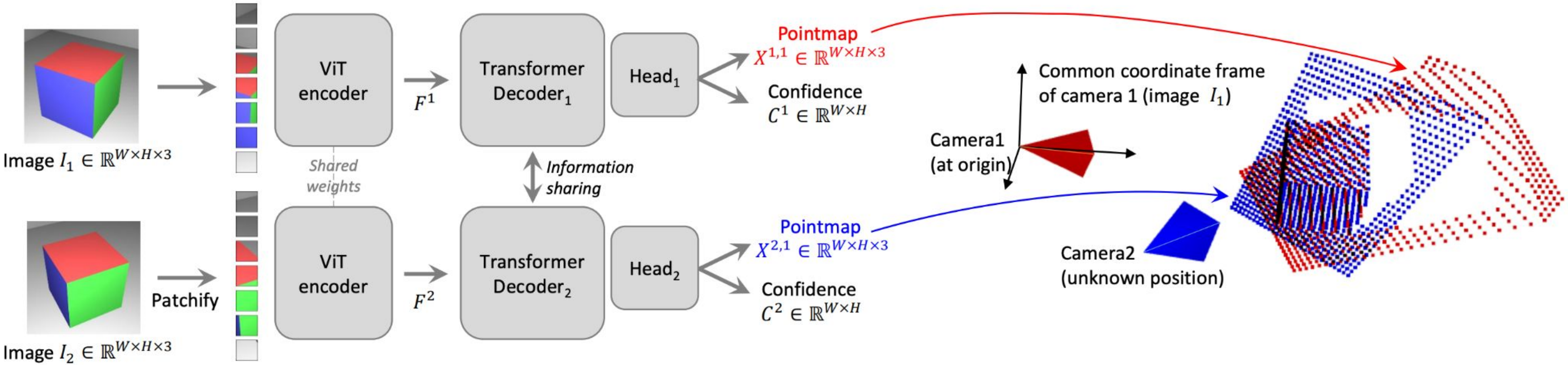


Follow-up:
MASt3R [2],
MASt3R-SfM [3]

[2] Leroy et al., "Grounding image matching in 3d with mast3r", arXiv 2024
[3] Duisterhof et al., "MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion", 3DV 2025

3D Reconstruction Networks - DUSt3R [1]

[1] Wang et al., “Dust3r: Geometric 3d vision made easy”, CVPR 2024



Follow-up:
MASt3R [2],
MASt3R-SfM [3]

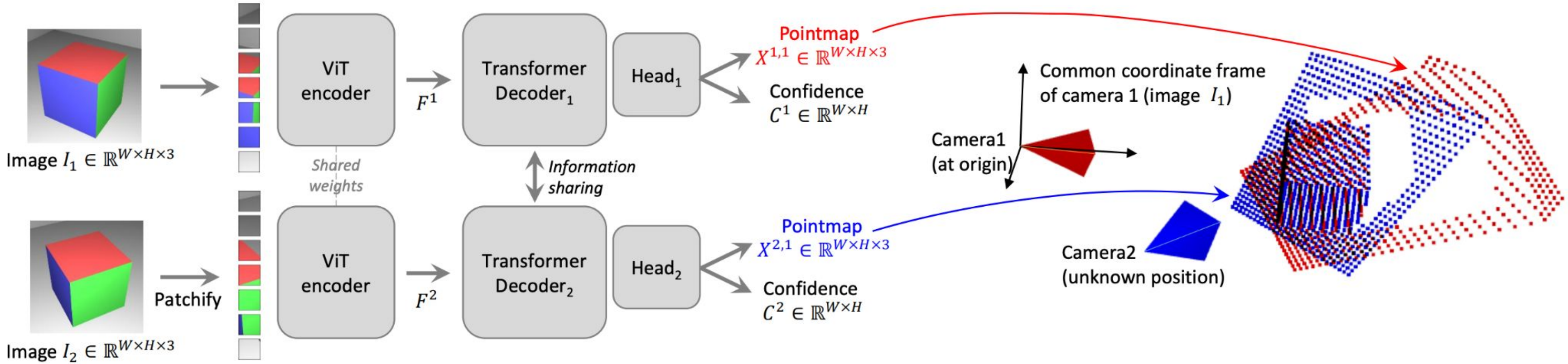
Multiple Views:
MV-DUSt3R+ [4],
VGGT [5]

[2] Leroy et al., “Grounding image matching in 3d with mast3r”, arXiv 2024
[3] Duisterhof et al., “MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion”, 3DV 2025

[4] Tang et al., “MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds”, CVPR 2025
[5] Wang et al., “VGGT: Visual geometry grounded transformer”, CVPR 2025

3D Reconstruction Networks - DUSt3R [1]

[1] Wang et al., “Dust3r: Geometric 3d vision made easy”, CVPR 2024



Follow-up:
MASt3R [2],
MASt3R-SfM [3]

Multiple Views:
MV-DUSt3R+ [4],
VGGT [5]

Adapt to downstream:
MASt3R-SLAM [6],
InstantSplat [7],

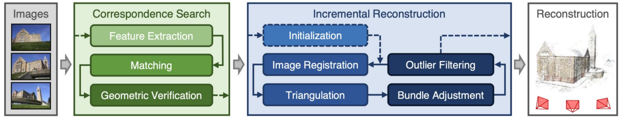
[2] Leroy et al., “Grounding image matching in 3d with mast3r”, arXiv 2024
[3] Duisterhof et al., “MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion”, 3DV 2025

[4] Tang et al., “MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds”, CVPR 2025
[5] Wang et al., “VGGT: Visual geometry grounded transformer”, CVPR 2025

[6] Murai et al., “MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors”, CVPR 2025
[7] Fan et al., “InstantSplat: Sparse-View Gaussian Splatting in Seconds”, arXiv 2024

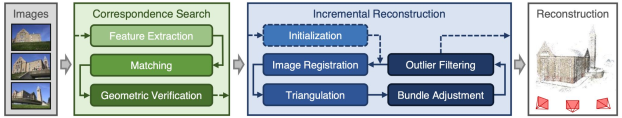
Reconstruction Evolution

Incremental SfM



Reconstruction Evolution

Incremental SfM



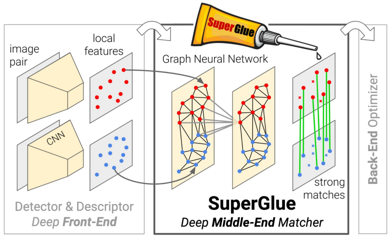
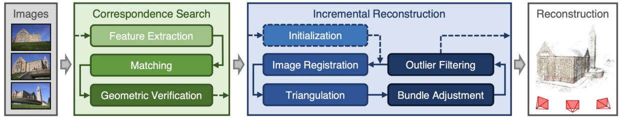


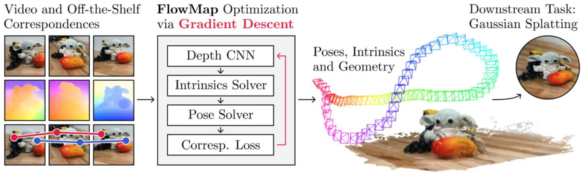
Image Matching

Reconstruction Evolution

Incremental SfM



End-to-End SfM



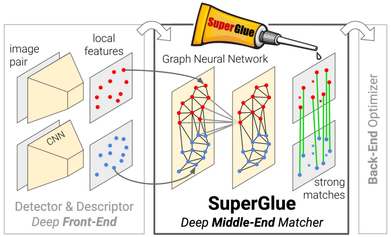
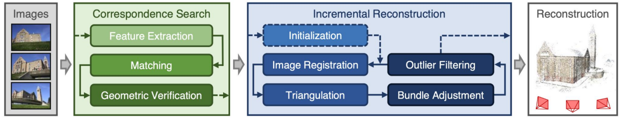


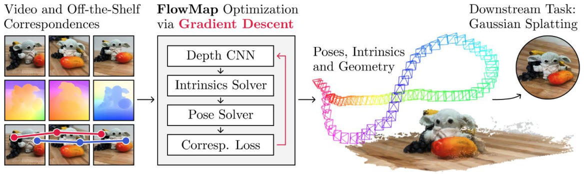
Image Matching

Reconstruction Evolution

Incremental SfM



End-to-End SfM



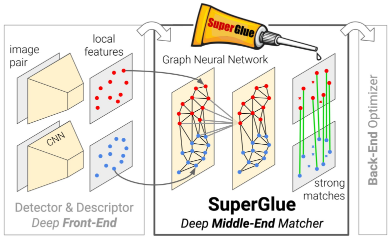
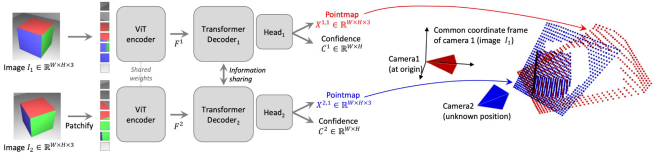


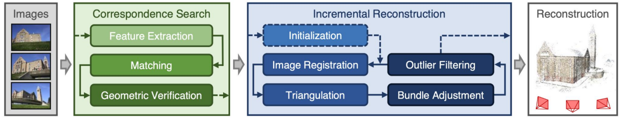
Image Matching



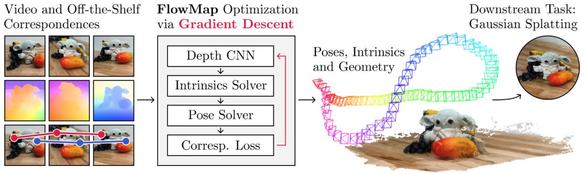
3D Reconstruction Networks

Reconstruction Evolution

Incremental SfM



End-to-End SfM



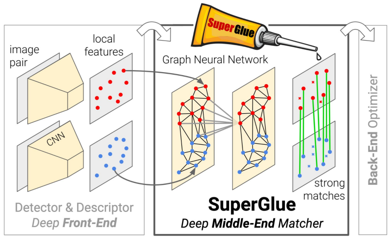
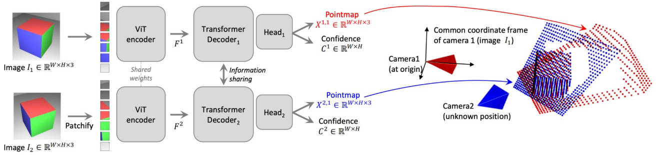


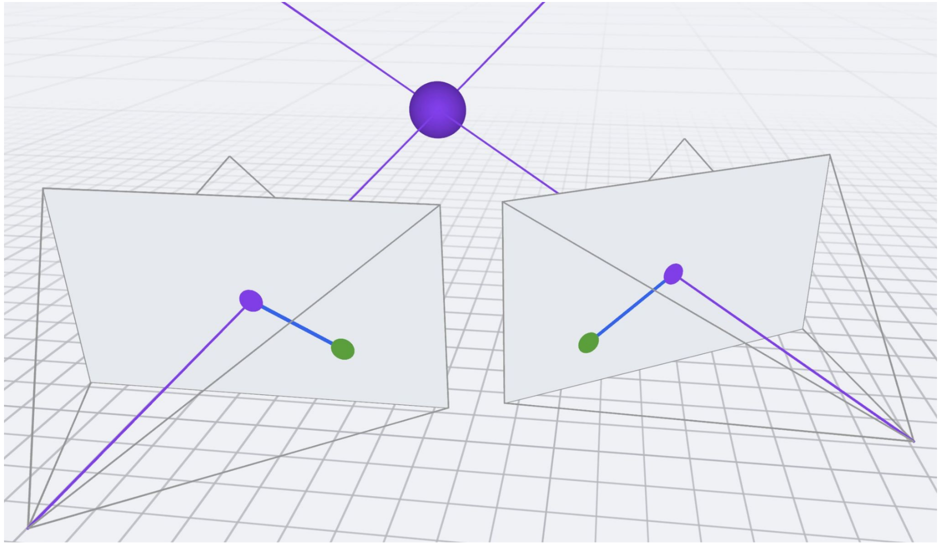
Image Matching



3D Reconstruction Networks

Preliminaries

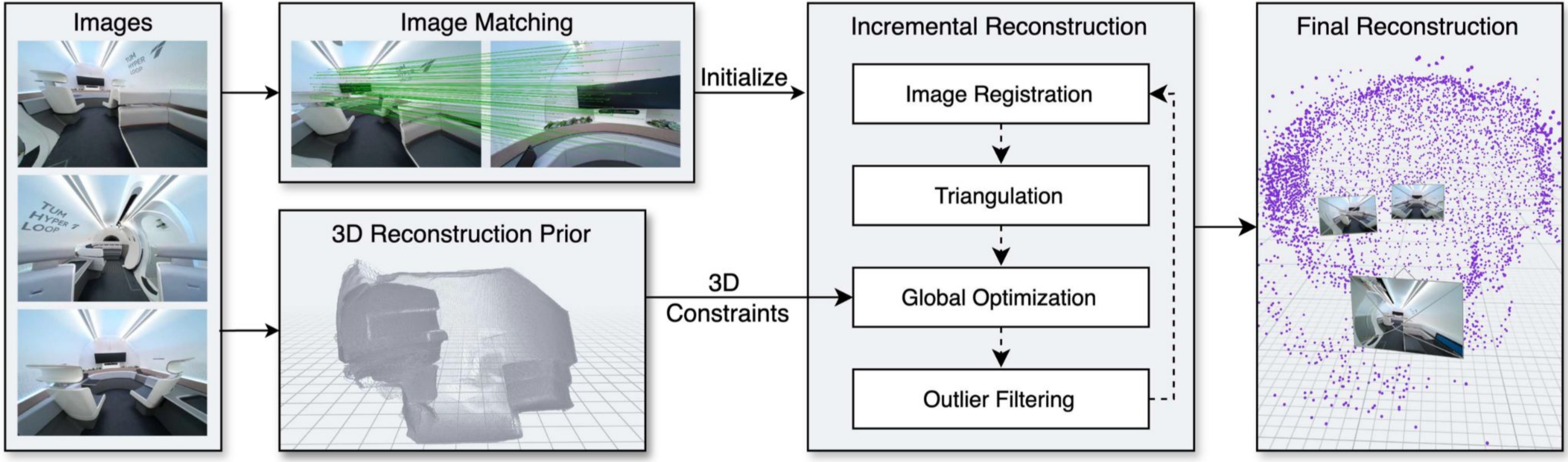
Bundle Adjustment



$$E_{\mathrm{BA}} = \sum_{i=1}^{N} \sum_{k=1}^{M} \| \boxed{y_{i,k}} - \pi(K_i, T_i, \boxed{x_k}) \|^2$$

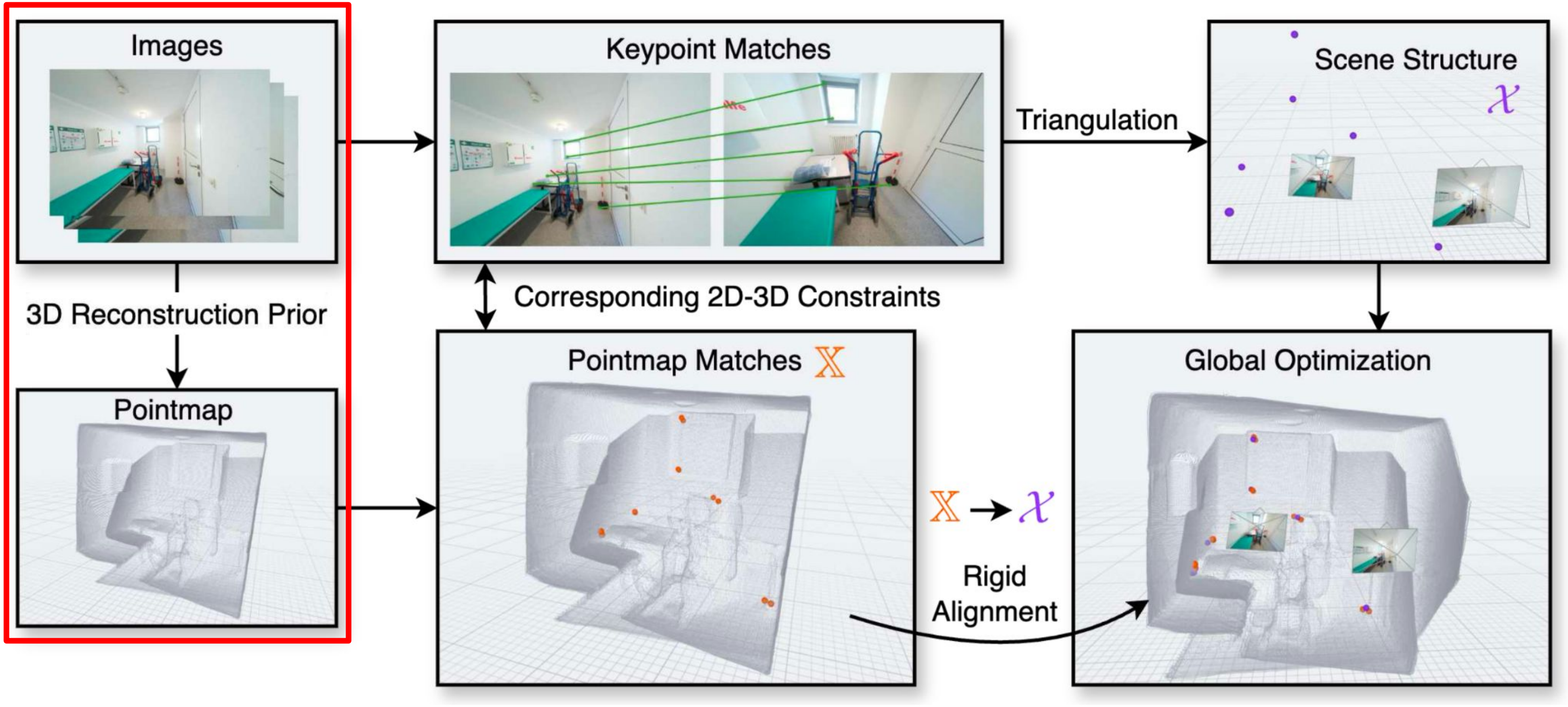
Method

Pipeline

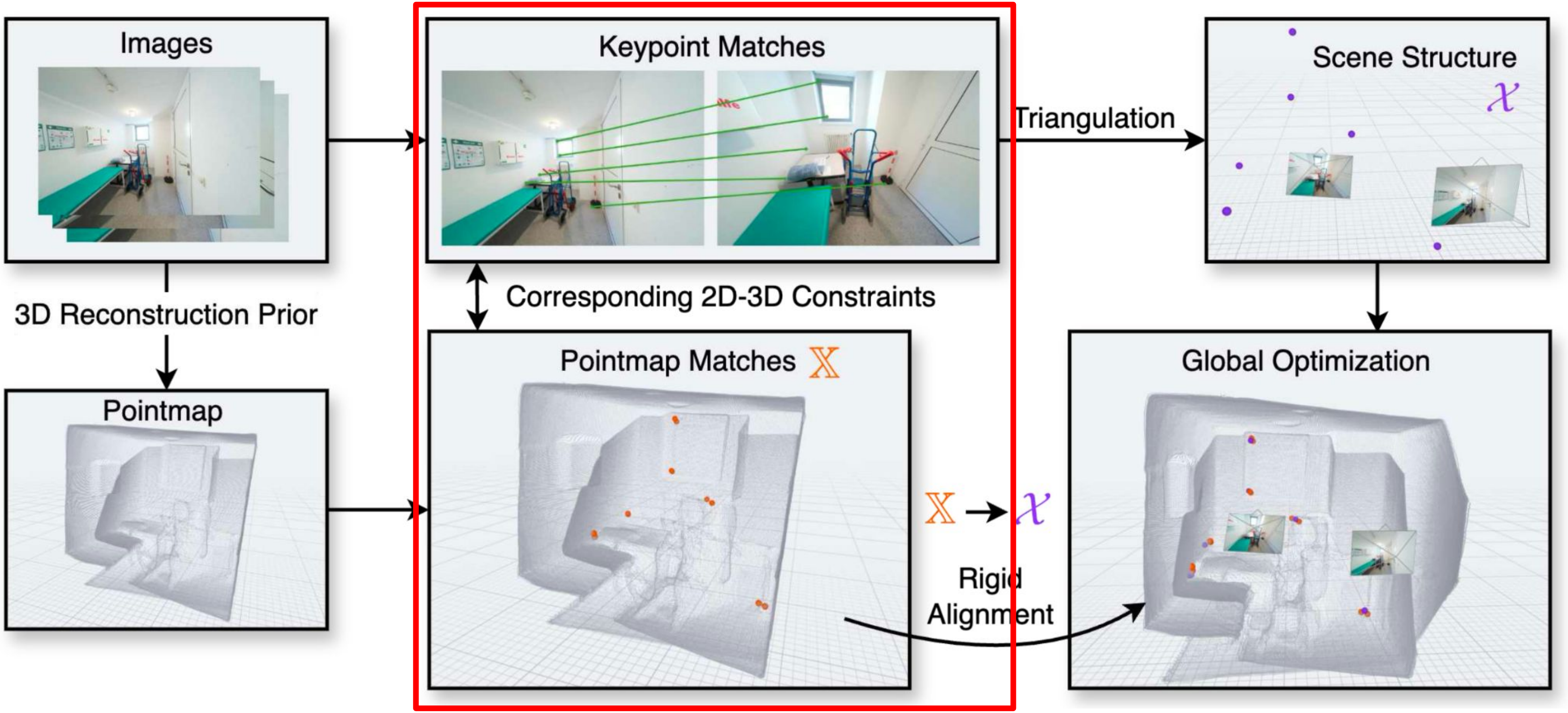


Images
Image Matching
Initialize
3D Reconstruction Prior
3D
Constraints
Incremental Reconstruction
Image Registration
Triangulation
Global Optimization
Outlier Filtering
Final Reconstruction

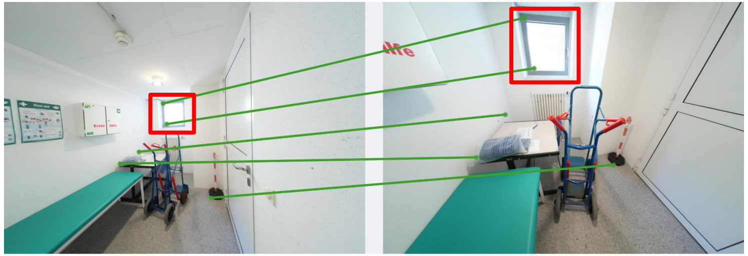
Method Overview

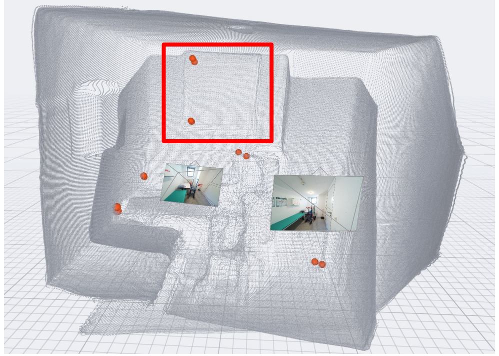


Method Overview

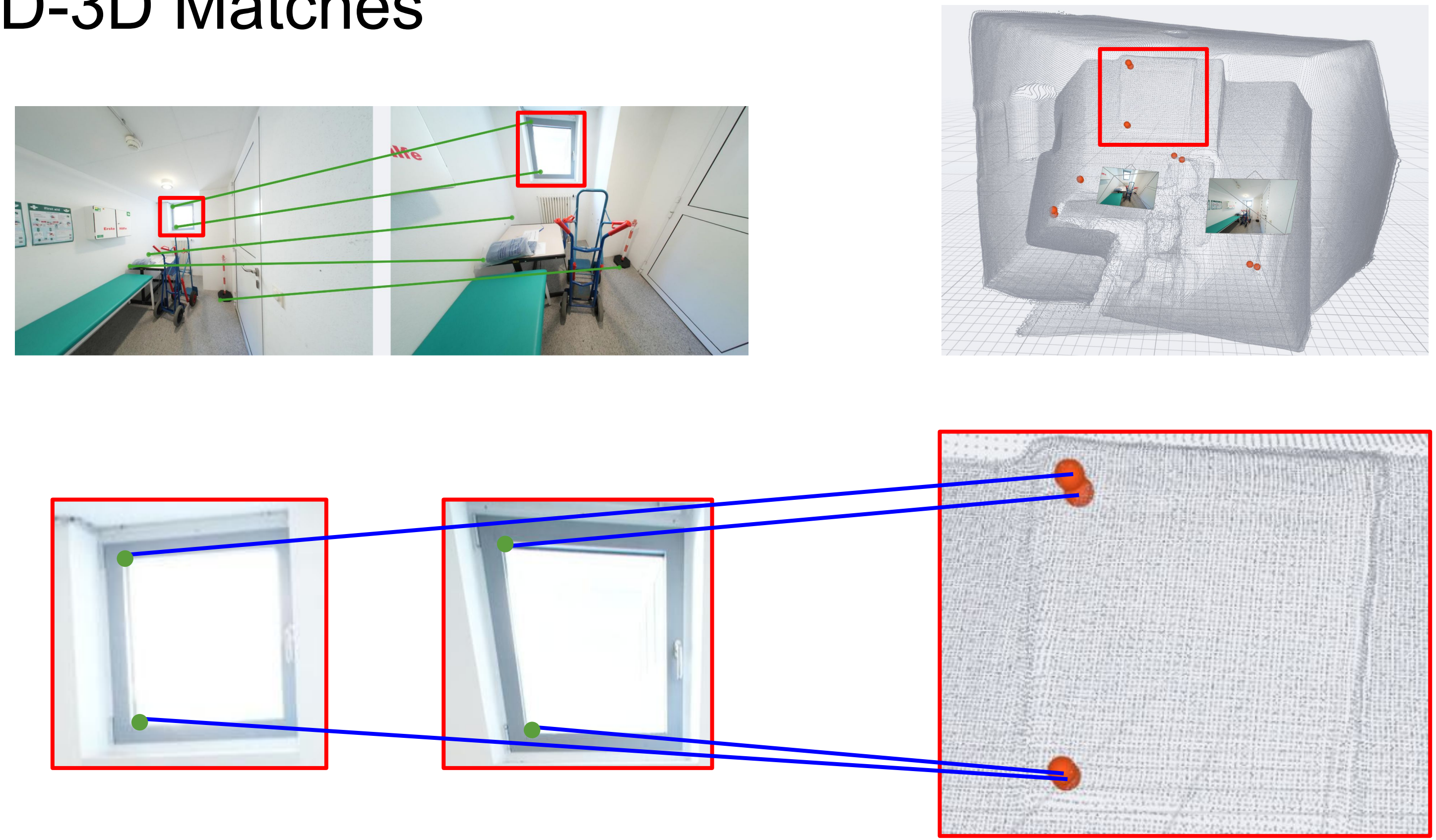


2D-3D Matches

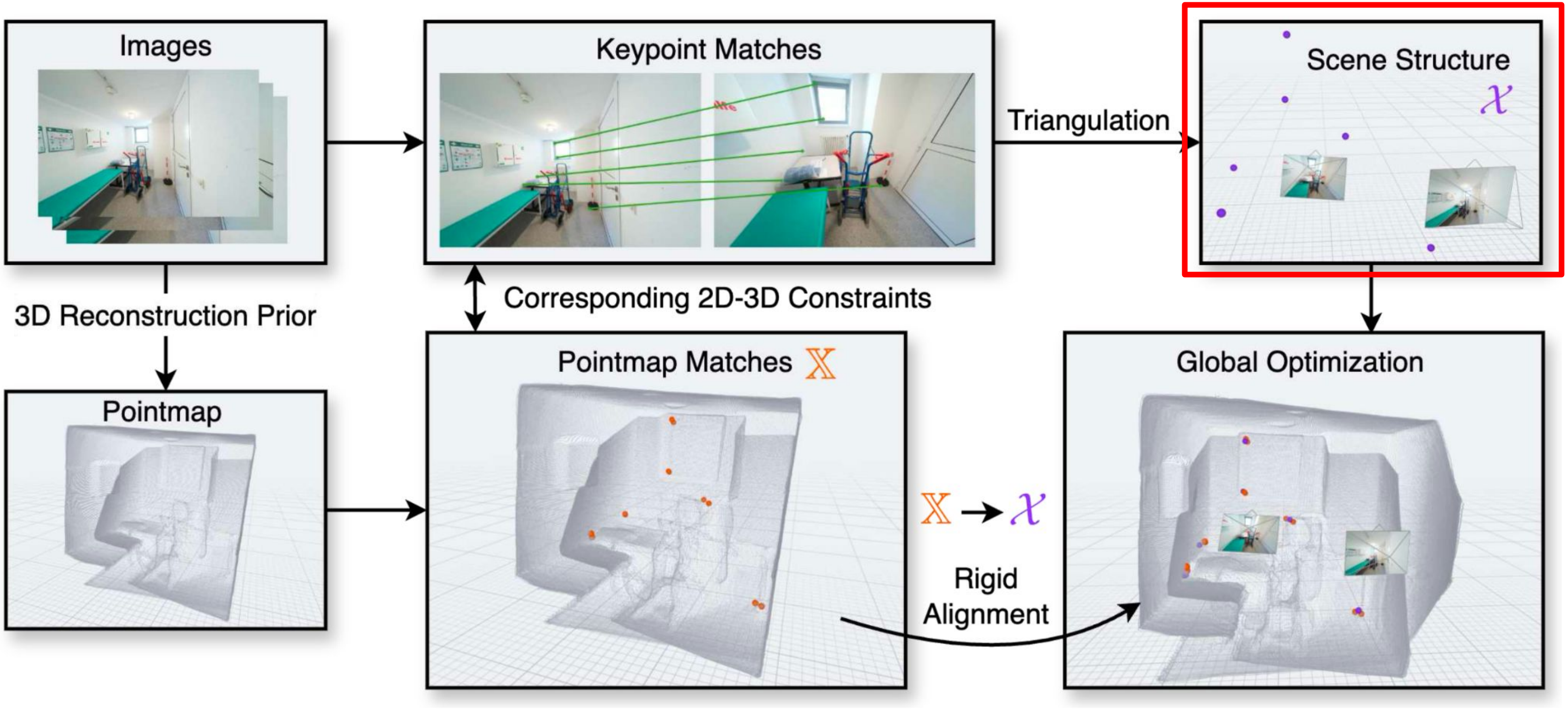




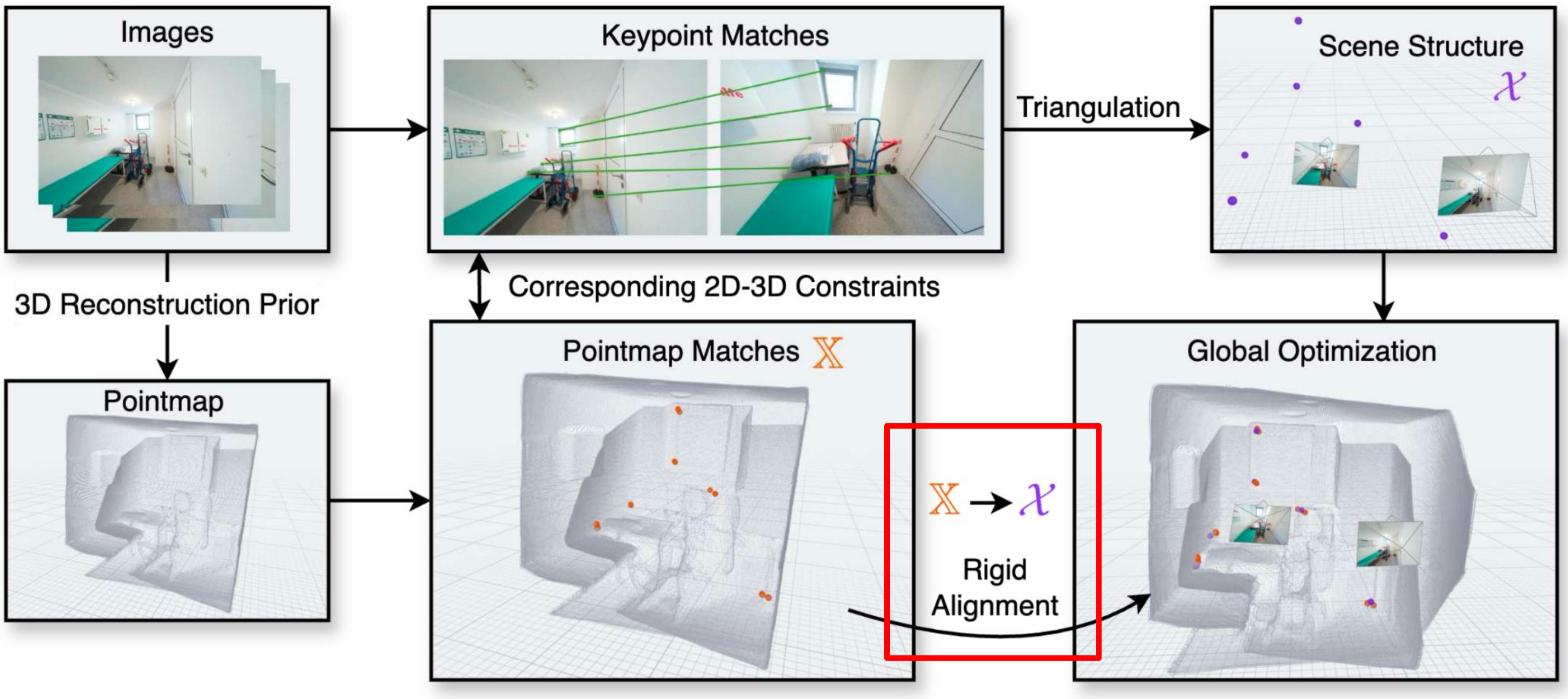
2D-3D Matches



Method Overview

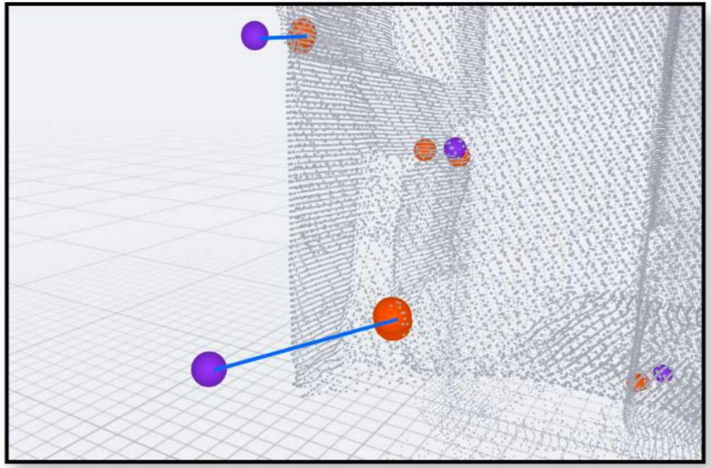


Method Overview



Global Optimization

Point-to-Point Error



$$\min_{\mathcal{X},\, T} \| \mathcal{X} - T(\mathbb{X}) \|$$

Intuitively: Make Scene Structure “agree” with 3D Reconstruction Networks

More Formally

$$E_{\mathrm{P2P}} = \sum_{e \in \mathcal{E}} \sum_{l \in \{i,j\}} \sum_{k=1}^{M_e} \| x_k - s_e T_e(x_k^{l,e}) \|^2$$

More Formally

scene point

pointmap point

$$E_{\mathrm{P2P}} = \sum_{e \in \mathcal{E}} \sum_{l \in \{i,j\}} \sum_{k=1}^{M_e} \| \boxed{x_k} - s_e T_e \boxed{(x_k^{l,e})} \|^2$$

More Formally

$$E_{\mathrm{P2P}} = \sum_{e \in \mathcal{E}} \sum_{l \in \{i,j\}} \sum_{k=1}^{M_e} \| \overbrace{\boxed{x_k}}^{\text{scene point}} - \underbrace{\boxed{s_e T_e}}_{\substack{\text{rigid transformation} \\ \text{(+ scale)}}} \overbrace{\boxed{(x_k^{l,e})}}^{\text{pointmap point}} \|^2$$

More Formally

scene point

pointmap point

$$E_{\mathrm{P2P}} = \underbrace{\sum_{e \in \mathcal{E}} \sum_{l \in \{i,j\}}}_{\text{for all pairwise pointmaps}} \sum_{k=1}^{M_e} \| \boxed{x_k} - \boxed{s_e T_e} \boxed{(x_k^{l,e})} \|^2$$

rigid transformation
(+ scale)

for all pairwise
pointmaps

More Formally

scene point

pointmap point

$$E_{\text{P2P}} = \boxed{\sum_{e\in\mathcal{E}} \sum_{l\in\{i,j\}}} \boxed{\sum_{k=1}^{M_e}} \| \boxed{x_k} - \boxed{s_e T_e} \boxed{(x_k^{l,e})} \|^2$$

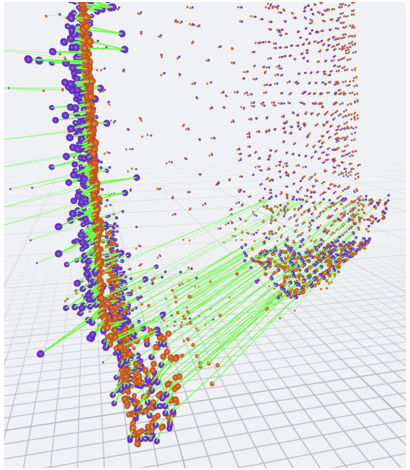
rigid transformation
(+ scale)

for all pairwise
pointmaps

for all matches

More Formally

$$E_{\mathrm{P2P}} = \sum_{e \in \mathcal{E}} \sum_{l \in \{i,j\}} \sum_{k=1}^{M_e^{*}} \| x_k - s_e T_e(x_k^{l,e}) \|^2$$



-> Rigid Alignment (RANSAC) + only minimize for inliers

More Formally

$$E_{\mathrm{P2P}} = \sum_{e \in \mathcal{E}} \sum_{l \in \{i,j\}} \sum_{k=1}^{M_e^*} \boxed{c_k^{l,e}} \| x_k - s_e T_e(x_k^{l,e}) \|^2$$

pointmap confidence -> downweight impact of inaccurate pointmaps

Global Optimization

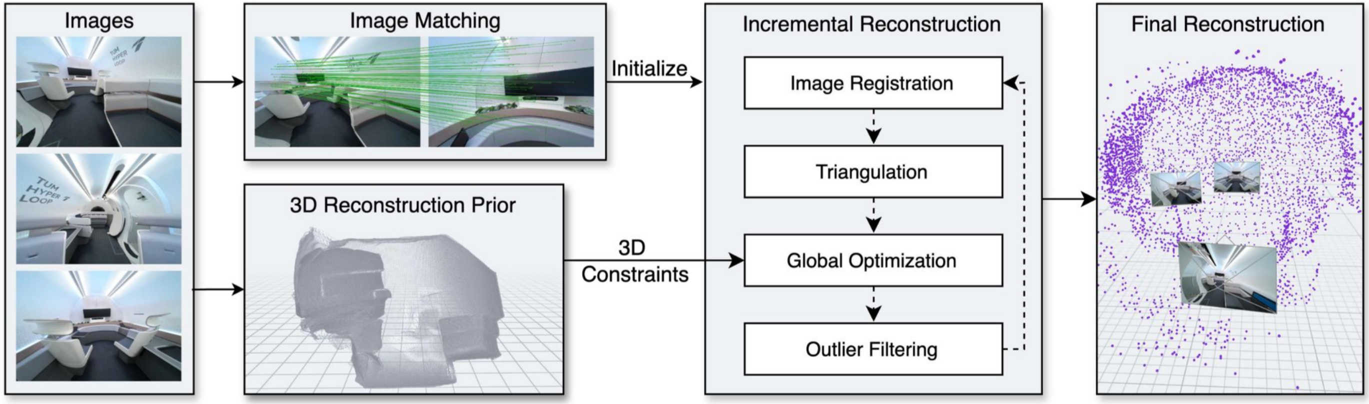
$$\mathcal{X}^*, \mathcal{H}^* = \underset{\mathcal{X}, \mathcal{H}, \mathcal{T}}{\arg\min} \left(E_{BA} + \beta E_{P2P} \right)$$

Rigid Transformations

Scene Cloud

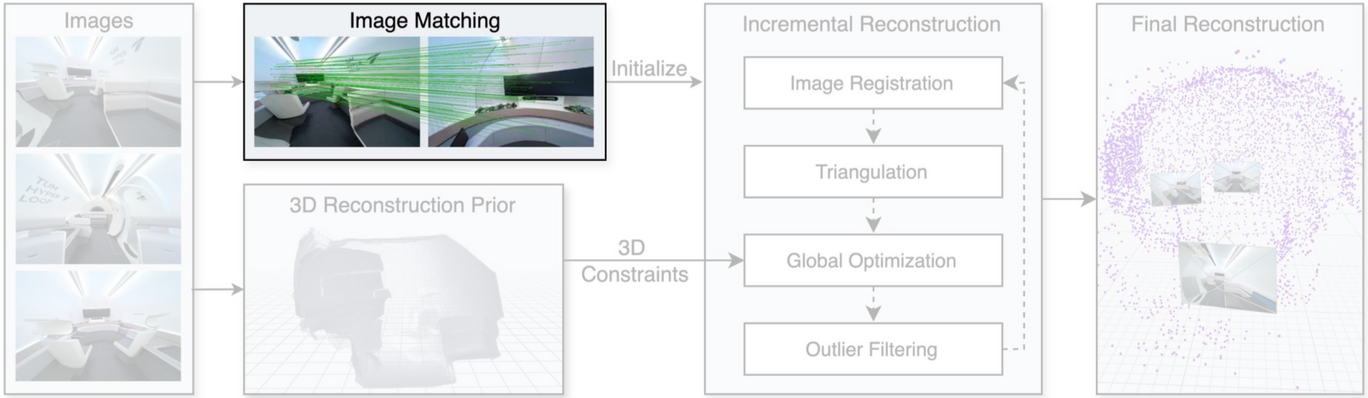
Camera Params

Implementation Details



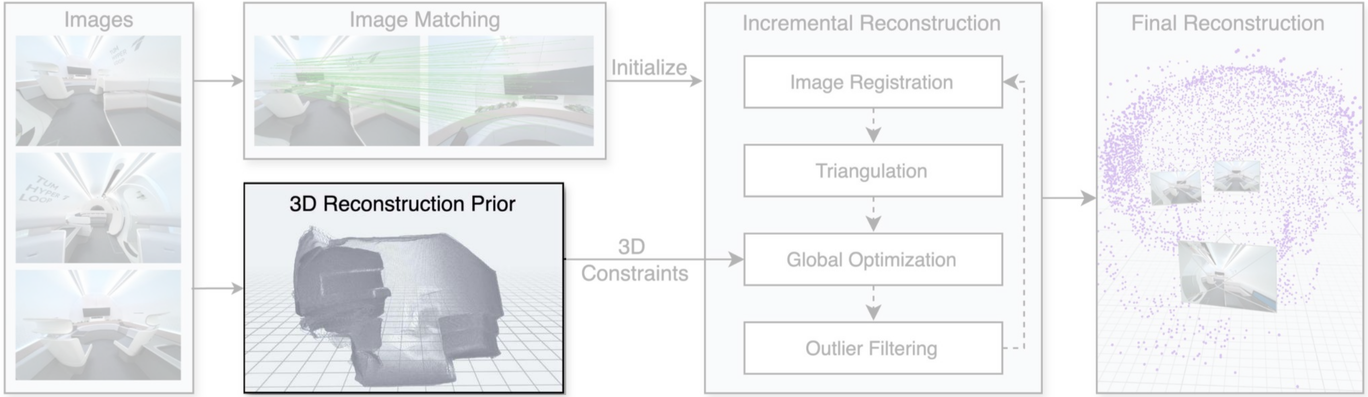
- implemented with opencv & torch

Implementation Details - Image Matching

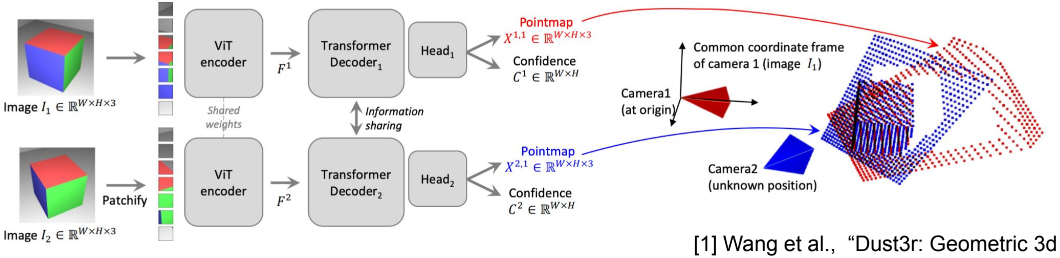


- Feature Extraction + Matching: MASt3R (limit to 256 matches)
- Geometric Verification: Essential Matrix + RANSAC

Implementation Details - 3D Reconstruction Prior

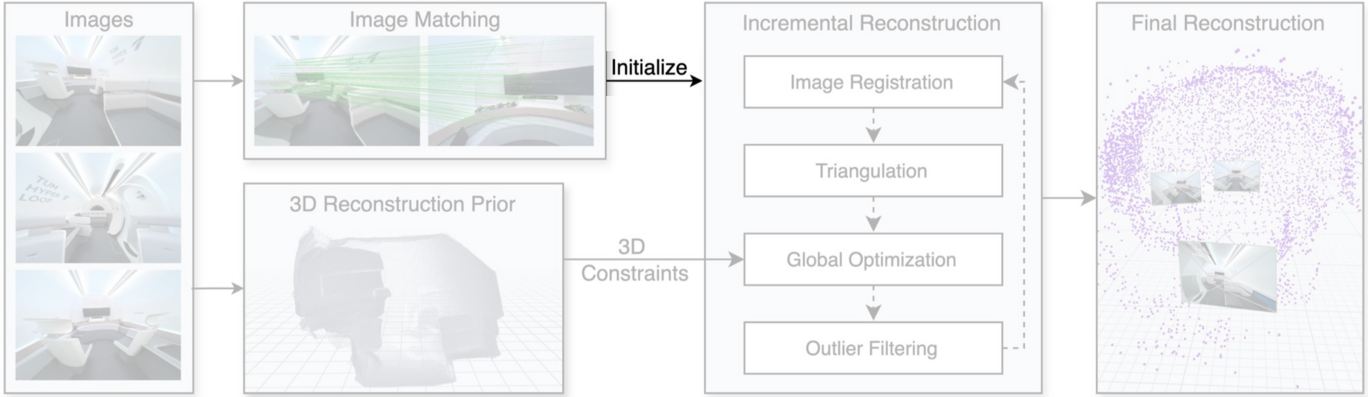


- DUSt3R 512x512 input res + DPT [1]



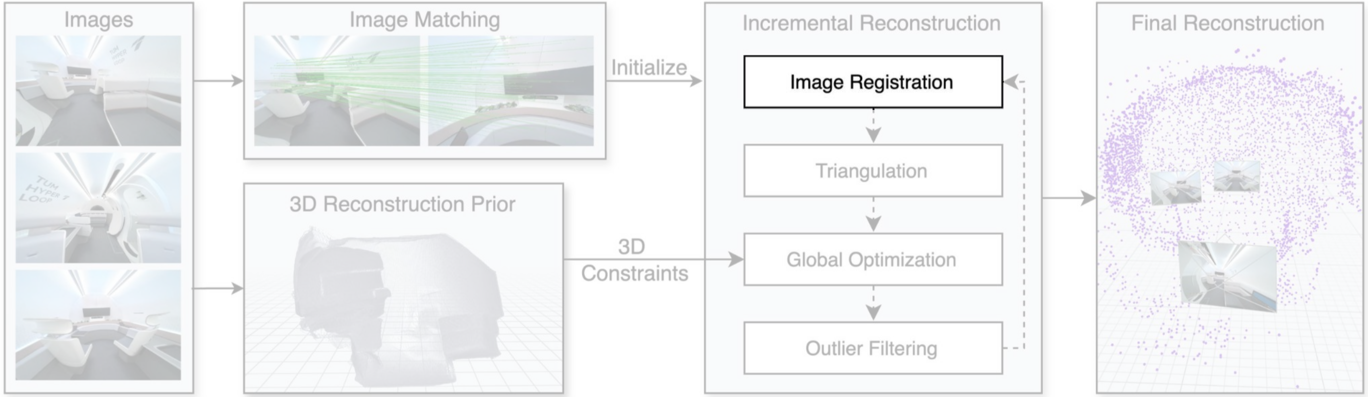
[1] Wang et al., “Dust3r: Geometric 3d vision made easy”, CVPR 2024

Implementation Details - Initialization



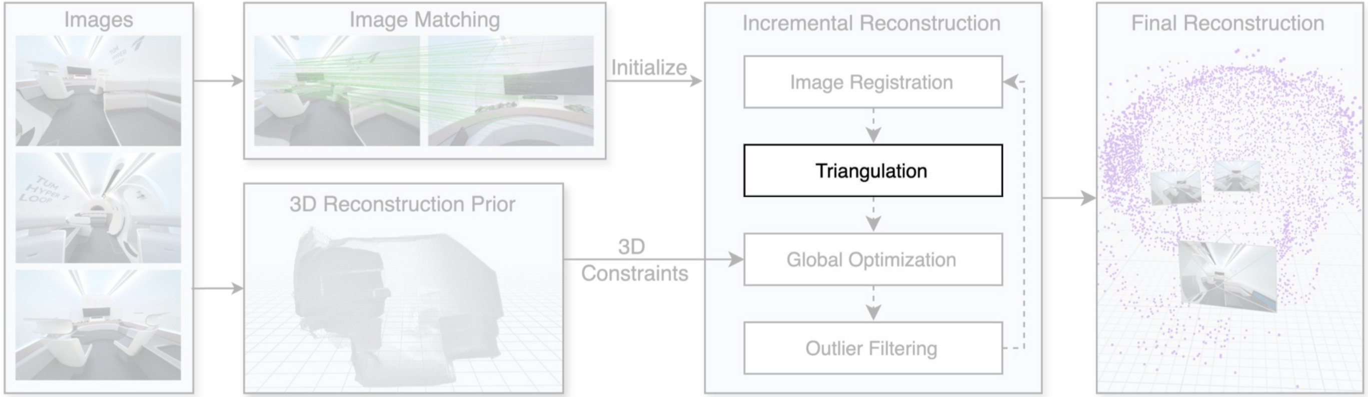
- Select initial pair based on #Matches and median triangulation angle

Implementation Details - Image Registration



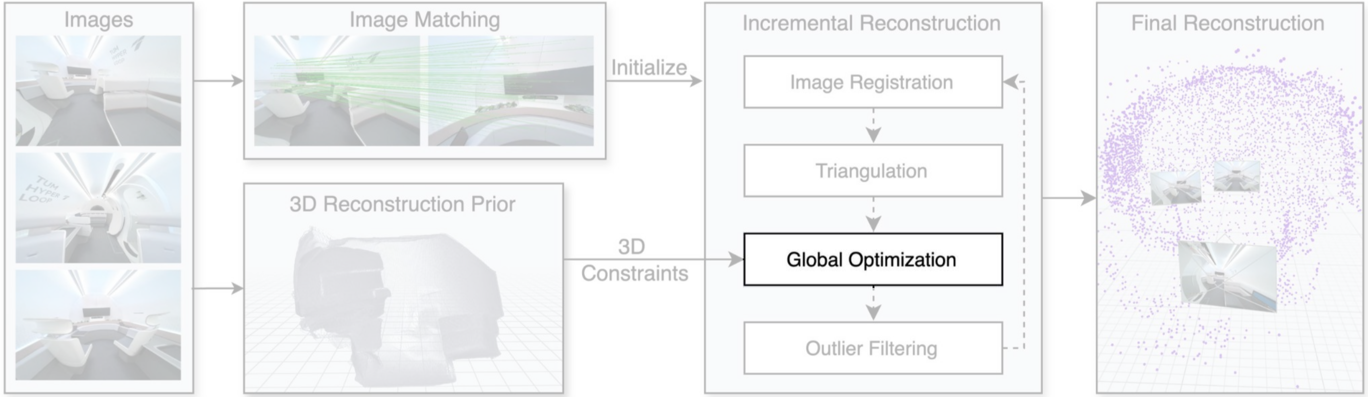
- Next Best View: #Visible Points
- Registration: PnP + RANSAC

Implementation Details - Triangulation



- Multi-View Triangulation (using DLT Method)
- Reject points with high reprojection error or low triangulation angle

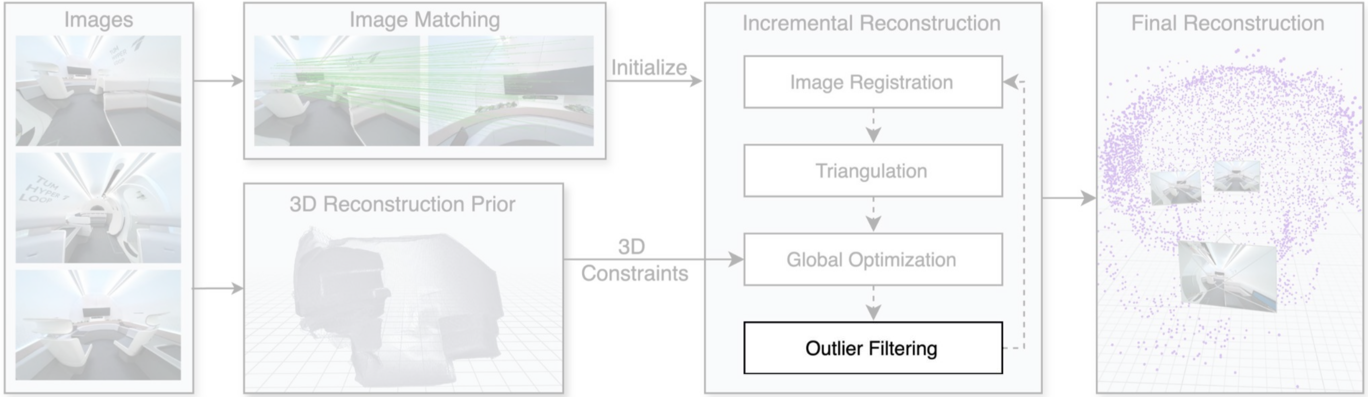
Implementation Details - Global Optimization



1. Pairwise RANSAC Alignment to Global Scene (use as initial parameters)
2. Remove outliers from energy
3. Minimize (GD + Linesearch)

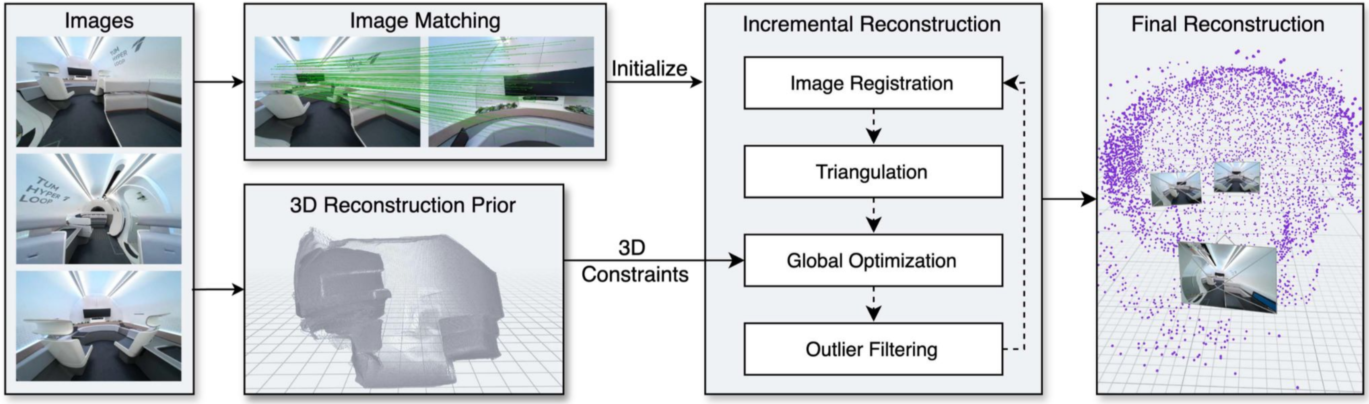
$$\mathcal{X}^*, \mathcal{H}^* = \underset{\mathcal{X},\mathcal{H},\mathcal{T}}{\arg\min}(E_{BA} + \beta E_{P2P})$$

Implementation Details - Outlier Filtering

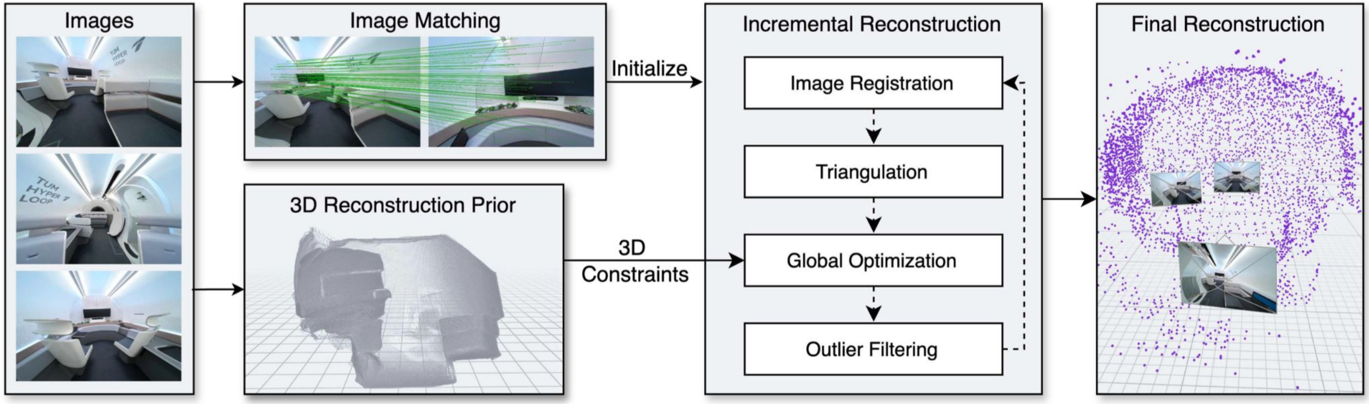


- High Reprojection Error
- Low Triangulation Angle

Implementation Details



Implementation Details - TEMPLATE SLIDE



- TEMPLATE SLIDE

Experiments

Experimental Setup - Metrics

Methods:

- Baseline
- Baseline+Ours

—

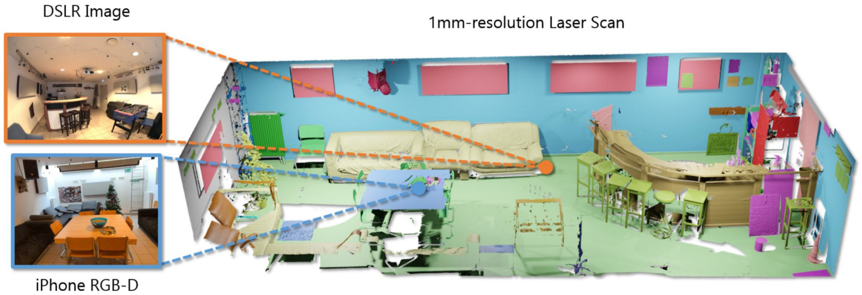
- DUSt3R + GO [1]
- VGGT [2]
- MASt3R-SfM [3]

Metrics:

- Average Translation Error (ATE)
- AUC@30
- Registration Rate

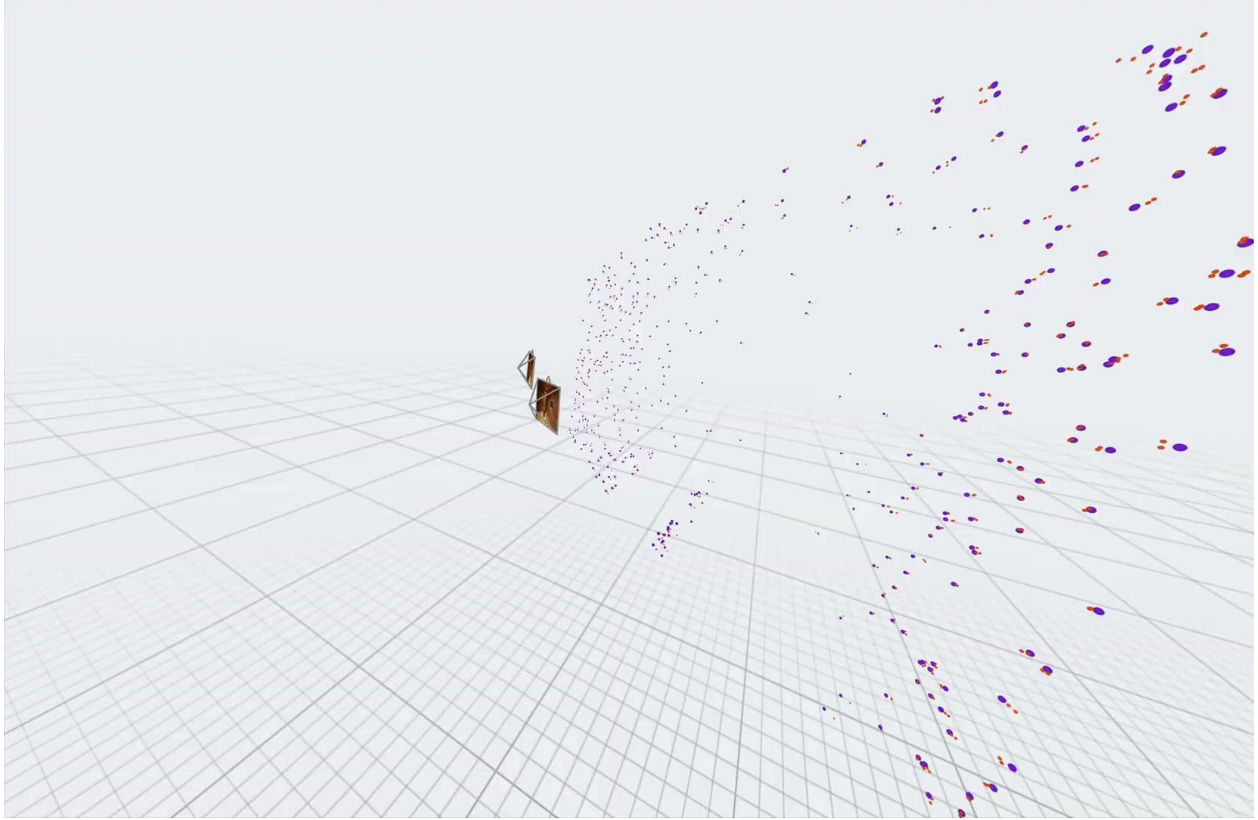
Data:

- ScanNet++ [4] **v2** scenes
- pseudo-GT through COLMAP

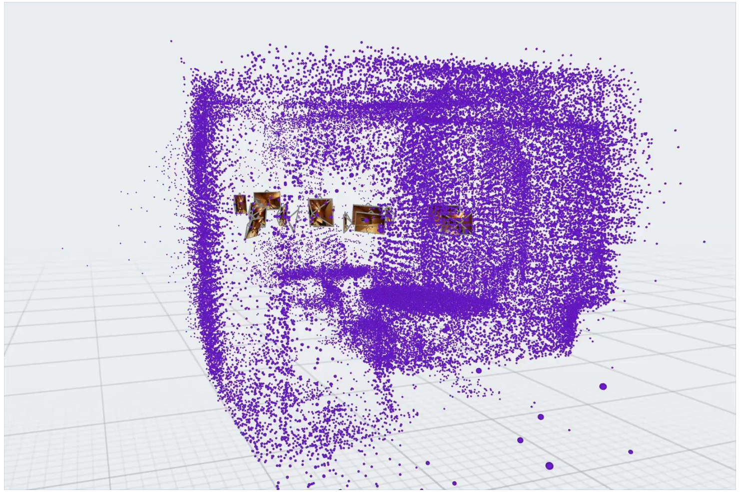


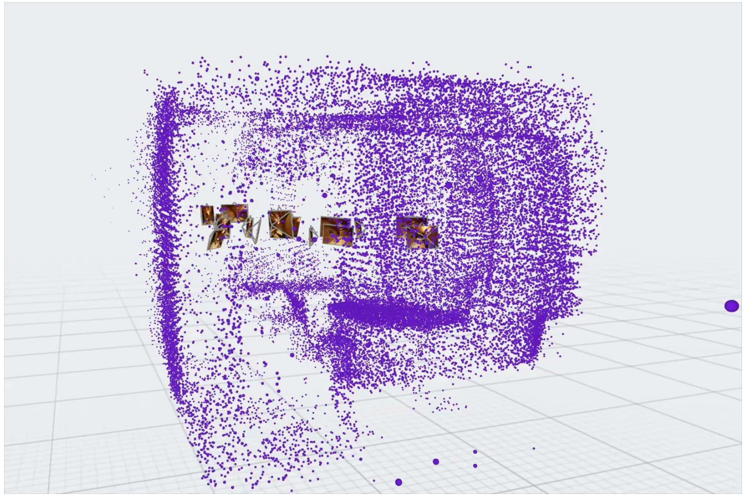
[1] Wang et al., "Dust3r: Geometric 3d vision made easy", CVPR 2024
[2] Wang et al., "VGGT: Visual geometry grounded transformer", CVPR 2025
[3] Duisterhof et al., "MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion", 3DV 2025
[4] Yeshwanth & Liu et al., ScanNet++: A high-fidelity dataset of 3d indoor scenes, ICCV 2023

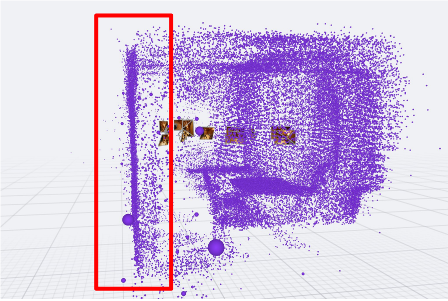
Visualization of Reconstruction Process



Visual Comparison

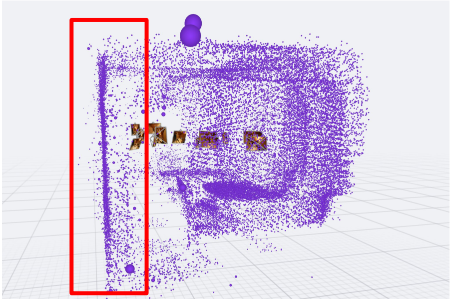






Baseline

Ours



*baseline has more scene points in general

Main Results

Method	15 Images			20 Images			25 Images		
	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑
Baseline	**0.0181**	82.4	97.1	0.0117	86.6	98.0	0.0107	86.7	99.3
Baseline+Ours	0.0190	**83.5**	96.9	**0.0090**	**88.3**	98.7	**0.0074**	**90.8**	98.6

Table 1. Camera pose estimation on ScanNet++ [31] with varying view counts (15, 20, 25). ATE (↓), AUC@30 (↑), and registration rate (↑). Metrics averaged over 30 scenes. *Feed-forward pose regression without further optimization.

Main Results

Method	15 Images			20 Images			25 Images		
	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑
Baseline	**0.0181**	82.4	97.1	0.0117	86.6	98.0	0.0107	86.7	99.3
Baseline+Ours	0.0190	**83.5**	96.9	**0.0090**	**88.3**	98.7	**0.0074**	**90.8**	98.6

Table 1. Camera pose estimation on ScanNet++ [31] with varying view counts (15, 20, 25). ATE (↓), AUC@30 (↑), and registration rate (↑). Metrics averaged over 30 scenes. *Feed-forward pose regression without further optimization.

Main Results

Method	15 Images			20 Images			25 Images		
	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑
Baseline	**0.0181**	82.4	97.1	0.0117	86.6	98.0	0.0107	86.7	99.3
Baseline+Ours	0.0190	**83.5**	96.9	**0.0090**	**88.3**	98.7	**0.0074**	**90.8**	98.6

Table 1. Camera pose estimation on ScanNet++ [31] with varying view counts (15, 20, 25). ATE (↓), AUC@30 (↑), and registration rate (↑). Metrics averaged over 30 scenes. *Feed-forward pose regression without further optimization.

Main Results

Method	15 Images			20 Images			25 Images		
	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑
Baseline	**0.0181**	82.4	97.1	0.0117	86.6	98.0	0.0107	86.7	99.3
Baseline+Ours	0.0190	**83.5**	96.9	**0.0090**	**88.3**	98.7	**0.0074**	**90.8**	98.6

Table 1. Camera pose estimation on ScanNet++ [31] with varying view counts (15, 20, 25). ATE (↓), AUC@30 (↑), and registration rate (↑). Metrics averaged over 30 scenes. *Feed-forward pose regression without further optimization.

Main Results

Method	15 Images			20 Images			25 Images		
	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑	ATE ↓	AUC@30 ↑	Reg. ↑
Baseline	**0.0181**	82.4	97.1	0.0117	86.6	98.0	0.0107	86.7	99.3
Baseline+Ours	0.0190	**83.5**	96.9	**0.0090**	**88.3**	98.7	**0.0074**	**90.8**	98.6
DUSt3R+GO	0.0234	80.8	**100**	0.0147	84.7	**100**	0.0134	85.2	**100**
VGGT*	0.0240	69.9	**100**	0.0192	71.4	**100**	0.0179	71.5	**100**
MASt3R-SfM	0.0211	76.3	**100**	0.0133	78.8	**100**	0.0118	78.8	**100**

Table 1. Camera pose estimation on ScanNet++ [31] with varying view counts (15, 20, 25). ATE (↓), AUC@30 (↑), and registration rate (↑). Metrics averaged over 30 scenes. *Feed-forward pose regression without further optimization.

Ablations

Energy Design Choices

Method	ATE ↓	AUC@30 ↑	Reg. ↑	#Pts ↑
Baseline	0.0159	80.6	95.3	1204
+P2P	0.0736	54.0	74.0	795

Table 2. Ablation study on design choices for our energy formulation. Metrics are averaged over 15 images from 10 different scenes in ScanNet++ [31].

Energy Design Choices

Method	**ATE** $\downarrow$	**AUC@30** $\uparrow$	**Reg.** $\uparrow$	**#Pts** $\uparrow$
Baseline	0.0159	80.6	95.3	1204
+P2P	0.0736	54.0	74.0	795
+*Inliers only*	0.0166	82.6	94.0	1224

Table 2. Ablation study on design choices for our energy formulation. Metrics are averaged over 15 images from 10 different scenes in ScanNet++ [31].

Energy Design Choices

Method	ATE $\downarrow$	AUC@30 $\uparrow$	Reg. $\uparrow$	#Pts $\uparrow$
Baseline	0.0159	80.6	95.3	1204
+P2P	0.0736	54.0	74.0	795
+*Inliers only*	0.0166	82.6	94.0	1224
+*Conf. Weight*	**0.0138**	**84.9**	**98.0**	**1260**

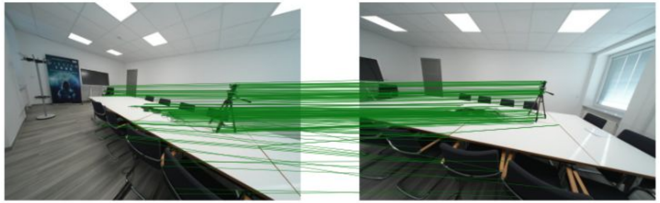
Table 2. Ablation study on design choices for our energy formulation. Metrics are averaged over 15 images from 10 different scenes in ScanNet++ [31].

Image Matching (2D Constraints)

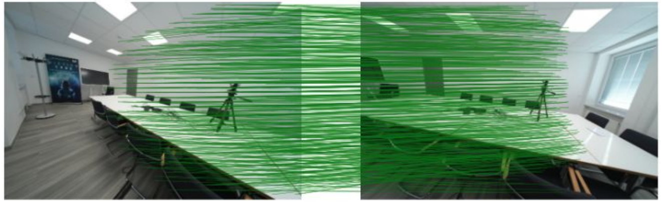
Matches	Method	ATE ↓	AUC@30 ↑	Reg. ↑
SIFT+NN	Baseline	0.0243	73.3	**64.0**
	+Ours	**0.0228**	**73.8**	**64.0**
MASt3R	Baseline	0.0159	80.6	95.3
	+Ours	**0.0138**	**84.9**	**98.0**

Table 3. Ablation study on different image matching methods (2D constraints). NN stands for nearest neighbor, MASt3R matches are computed using fast reciprocal matching [14]. Metrics are averaged over 10 ScanNet++ [31] scenes, each with 15 images.

SIFT matches



MASt3R matches



3D Reconstruction Prior (3D Constraints)

3D Reconstruction Prior	ATE ↓	AUC@30 ↑	Reg. ↑
Baseline (No Prior)	0.0159	80.6	95.3
DUSt3R	0.0138	**84.9**	**98.0**
VGGT	0.0137	82.61	96.7
VGGT-MV	**0.0110**	84.06	97.3

Table 4. Ablation study on different 3D reconstruction priors. VGGT-MV extracts multi-view pointmaps instead of pairwise ones. Metrics are averaged over 10 ScanNet++ [31] scenes, each with 15 images.

3D Reconstruction Prior (3D Constraints)

3D Reconstruction Prior	**ATE** ↓	**AUC@30** ↑	**Reg.** ↑
Baseline (No Prior)	0.0159	80.6	95.3
DUSt3R	0.0138	**84.9**	**98.0**
VGGT	0.0137	82.61	96.7
VGGT-MV	**0.0110**	84.06	97.3

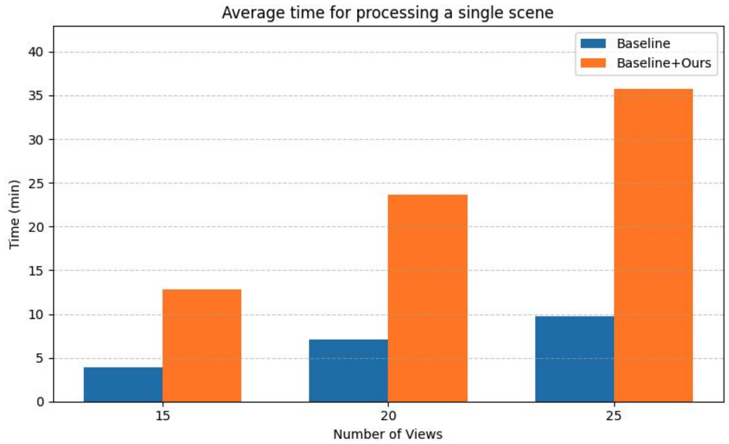
Table 4. Ablation study on different 3D reconstruction priors. VGGT-MV extracts multi-view pointmaps instead of pairwise ones. Metrics are averaged over 10 ScanNet++ [31] scenes, each with 15 images.

Limitations & Future Work

Scalability

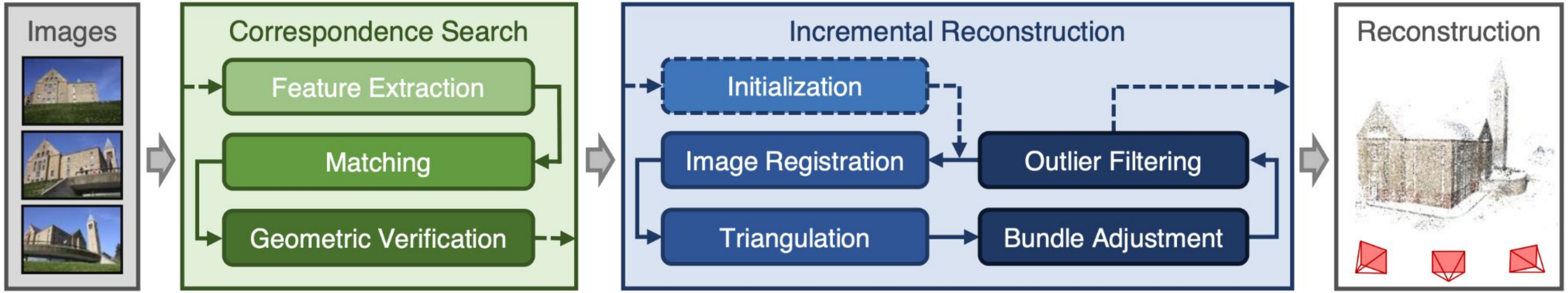
N Images -> up to [N choose 2] $\binom{N}{2}$
pairwise pointmaps

-> Multi-View Methods
-> Merging pairwise pointmaps
during scene alignment



Integrate into other parts of the pipeline

3D constraints **only** valid in Global Optimization, rest of pipeline relies **solely** on 2D keypoint matches



[1] Schönberger and Frahm, “Structure-from-motion revisited”, CVPR 2016

Conclusion

Visual Computing &
Artificial Intelligence
Prof. Matthias Nießner

Revisiting Structure from Motion with 3D Reconstruction Priors

Guided Research WS24/25
Daniel Korth
Advisor: Prof. Matthias Nießner

30.05.2025